

文章编号: 2095-4980(2019)05-0866-05

机器学习自动识别小鼠异位性心搏

何 密^{1a}, 粘永健^{1a}, 张 芸², 林哲宇³, 胡厚源^{1b}

(1.陆军军医大学(第三军医大学) a.生物医学工程与影像医学系; b.西南医院心内科, 重庆 400038;
2.解放军 75841 部队, 湖南 长沙 410007; 3.兰州药品仪器检验所, 甘肃 兰州 730050)

摘 要: 临床上检测异位性心搏, 对心血管疾病的早发现、早诊断、早治疗有重大意义。异位性心搏的自动识别能有效减轻人工识别的负担。采用 37 只小鼠各 10 min 的心电图(ECG)信号, 由 3 位专家完成异位性心搏的标定及数据库的建立。在此基础上, 使用 7 种机器学习方法, 组合脉冲排除滤波器和模板匹配算法的值, 对异位性心搏点进行自动识别。实验结果表明, 7 种机器学习的方法都能得到较好的预测性能(所有曲线下面积(AUC)大于 0.899), 其中集成学习 AdaBoost 的预测能力最好(AUC=0.940, 准确度、特异性均为 0.888)。

关键词: 异位性心搏; 脉冲排除滤波器; 模板匹配; 机器学习

中图分类号: TP311.5

文献标志码: A

doi: 10.11805/TKYDA201905.0866

Automatic recognition of mouse ectopic beats using machine learning

HE Mi^{1a}, NIAN Yongjian^{1a}, ZHANG Yun², LIN Zheyu³, HU Houyuan^{1b}

(1a.School of Biomedical Engineering and Imaging Medicine; 1b.Department of Cardiology of Southwest Hospital, Army Medical University(Third Military Medical University), Chongqing 400038, China; 2.No. 75841 Army of PLA, Changsha Hunan 410007, China; 3.Lanzhou Institute of Pharmaceutical Instruments, Lanzhou Gansu 730050, China)

Abstract: Clinical examination of ectopic beats is very important for early detection, diagnosis and treatment of cardiovascular diseases. Automatic recognition of ectopic beats can effectively reduce the burden of manual recognition. In this paper, the 10 minutes Electrocardiogram(ECG) signals of 37 mice were used for analysis. All ectopic beats were calibrated by 3 experts to establish the database. Using 7 machine learning methods, the ectopic beats were automatically identified by combining the values of Impulse Rejection Filter(IRF) and the template matching algorithm. The experimental results show that 7 machine learning methods can achieve good predictive performances(all Area Under Curve(AUC)>0.899), where the ensemble learning method—AdaBoost has the best predictive performance(AUC=0.940, sensitivity=sensitivity=0.888).

Keywords: ectopic beat; Impulse Rejection Filter; template matching; machine learning

随着社会的发展和城市居民生活方式、工作节奏与饮食规律的变化, 心脏疾病已成为现代社会的高发病症, 早期识别、正确诊断和及时治疗对降低心血管疾病的死亡率至关重要。心电图(ECG)有数百年应用的历史, 随着 ECG 仪器的发展日趋完善, 在心血管疾病的检测过程中具有不可替代的作用。其具有简单、快速、价格低廉的特点, 有着较高的临床应用价值^[1]。ECG 检查对患有心律失常病人的诊断有重要价值, 能直观体现患者心律失常的情况, 其中早搏性异位性心搏(包括房性早搏、室性早搏、房颤)是临床中最常见的心律失常类型^[2-3]。异位性心搏是由异常兴奋导致的不正常搏动^[4], 这些异常兴奋源自心房心室连接处或心室, 而不是窦房结。临床上检测异位性心搏, 对心血管疾病的早发现、早诊断、早治疗有重大意义。临床检测异位性心搏的方式有很多种, 如光电容积脉搏波法^[5]、测量血压信号^[6]和 ECG 法^[7]等, 其中 ECG 法是检测异位性心搏的金标准。

近年来, 相继涌现出多种 ECG 异位性心搏的自动识别算法。Mcnames 等提出一种非线性脉冲排除滤波器(IRF)^[8]识别 ECG 的异位性心搏, 但这种方法没有考虑 ECG 波形特征的改变, 从而导致部分错误识别。李峰等通过改进的二阶导数计算出数据点的斜率和运动趋势, 并结合 ECG 波形的实际物理特征, 实现了异位性心搏的自

收稿日期: 2018-10-22; 修回日期: 2019-01-05

基金项目: 中国博士后科学基金资助项目(2016M602991); 重庆市博士后科研资助项目(Xm2016120)

动识别^[9]。但室上性异位性心搏的波形特征与窦性的搏动相似,因此该算法在多源性的心律失常波形中,不能有效区别正常窦性搏动与室上性异位性心搏。LIU等提出一种改进的IRF方法,并采用简单的线性模型组合了改进的IRF方法和ECG波形特征模板匹配算法,取得了较好的预测效果^[10]。由于组合方法并没有将训练集和测试集分开,因此得到的结果不具有可信度。本文将小鼠心律失常模型的ECG数据分为训练集和测试集,提出基于机器学习的方法组合改进IRF和模板匹配算法,以提高异位性心搏识别的准确性。

1 实验方法

1.1 小鼠 ECG 数据库的建立

所用到的小鼠 ECG 信号由澳大利亚 Baker 心脏病和糖尿病研究所小鼠心脏病科研中心的实验人员采集和提供。实验的小鼠样本共有 3 类,分别为:野鼠(WT)12 只、 β_2 肾上腺素受体过表达的转基因小鼠(TG)18 只和 β_2 -TG 合并半乳糖凝集素-3(Gal-3)基因敲除的小鼠(TGKO)13 只^[11]。 β_2 -TG 鼠表现出心肌间质纤维化和爆发性室性心律失常,且随着年龄增大情况更严重。 β_2 肾上腺素受体在心肌上过度表达表现为心肌病的进行性发展,其特点是纤维化、心室重构和功能障碍、室性心律失常和过早死亡。Gal-3 是一种 β 半乳糖苷结合凝集素,结合多种细胞外和细胞内糖蛋白,从而调节其活性。Gal-3 的表达水平在病理条件下上调,临床上认为半 Gal-3 会促成多种疾病或病理过程(如癌症、炎症性疾病)以及器官纤维化。

一共采集 43 只小鼠的 ECG 信号,因干扰过强排除 5 只小鼠数据,其中 3 只 TG 小鼠,2 只 TGKO 小鼠;因数据长度记录不足,排除 1 只 TGKO 小鼠(仅 200 s 数据),最终剩下 37 只小鼠的 ECG 数据(2 只 WT 小鼠 ECG 信号、15 只 TG 小鼠 ECG 信号和 10 只 TGKO 小鼠 ECG 信号)用于算法的分析,如图 1 所示。

实验所用的小鼠月龄均为 9~10 个月,心率为 400~700 次/分,共记录了 2 个导联的 ECG 数据,分别是 Lead II 和 Chest Lead^[12]。Lead II 采集的是小鼠右前肢和左后肢皮下电极的信号;Chest Lead 的电极一端在左胸前大约肋骨下缘水平皮下,另一端在左前肢皮下。ECG 信号的采样率均为 2 000 Hz,每只时长大约为 10.5 min,截取 600 s 时长的 ECG 信号进行分析。

对于 β_2 肾上腺素受体过表达的转基因小鼠和其合并的 Gal-3 基因敲除的小鼠模型,异位性心搏的类别主要有室早、室上性早搏和室性心动过速。小鼠 ECG 的 R 波峰值点的自动检测可能会存在错误,在人工识别标定时,需要同时更正。专家识别小鼠 ECG 室早的标准是提早出现宽大畸形的 QRS 波群,QRS 波群前后无相关 P 波,代偿间歇多完全;小鼠 ECG 室上性早搏的标准是提早出现 P-QRS-T 波群,QRS 波群一般与窦性心律相似,P 波形态改变,代偿间歇不完全;室性心动过速的判别标准是连续出现 3 次及以上早搏的情形。

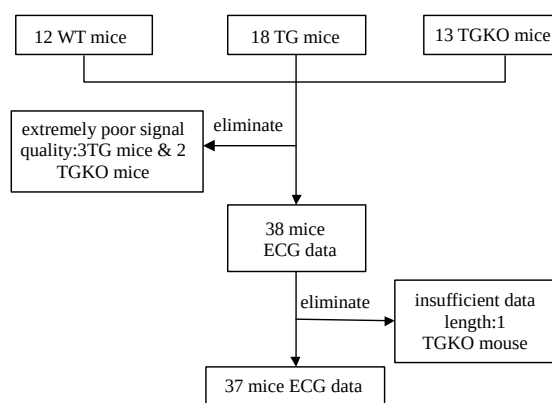


Fig.1 Flowchart of ECG data choice of mice
图 1 小鼠 ECG 数据选择的流程图

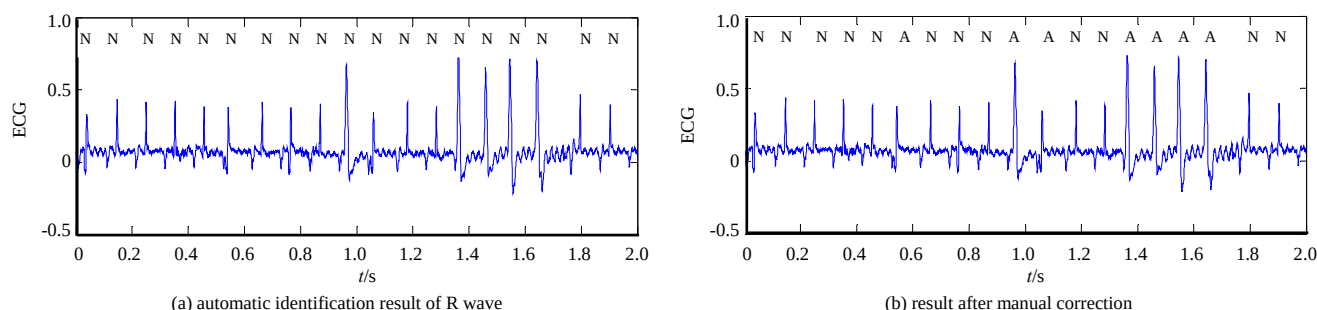


Fig.2 Calibration and correction of ectopic beats
图 2 异位性心搏的标定和纠正

人工标记的过程是找到异位性心搏的位置,将 N 改为 A。图 2 为异位性心搏点的标记和纠正,图中 N 为窦性心律 R 波的位置,图 2(a)为 R 波自动识别的结果,图 2(b)为人工标记后的结果。将 3 类小鼠 ECG 标记的结果合并,共得到 218 970 个样本,其中阳性样本 15 733 个,阴性样本 203 237 个。

1.2 机器学习方法

常用的机器学习算法有决策树算法、聚类算法、支持向量机(Support Vector Machine, SVM)算法、集成分类算法以及神经网络算法等。考虑到计算复杂度和样本量较大的情况,本文选用 SVM 算法、逻辑回归(Logistic Regression, LR)算法、K 近邻(k-Nearest Neighbor, KNN)算法、决策树(Decision Tree, DT)算法、朴素贝叶斯(Naive Bayes, NB)算法、二次判别分析(Quadratic Discriminant Analysis, QDA)算法、集成学习的 AdaBoost(Adaptive Boosting)算法^[13-17]7 种机器学习方法来自动识别异位性心搏。机器学习输入的特征是由 RRI(R-R-Interval)序列计算得到的 IRF 值和 ECG 波形计算得到的模板匹配值,输入的标签是人工识别的结果,正常搏动记为 0,异位性心搏则记为 1。

为消除指标间的量纲和取值范围差异的影响,减少数据的冗余性和不确定性,需要对输入的指标数据进行归一化处理,以保证计算结果的科学、准确性。本文采用 min-max 标准化方法归一化,将数值映射到[-1,1]区间,计算方法如下:

$$X^* = 2 \left(\frac{X - X_{\min}}{X_{\max} - X_{\min}} \right) - 1 \quad (1)$$

式中: $X_{\max}$ 为样本数据的最大值; $X_{\min}$ 为样本数据的最小值。

使用软件 Matlab2014a 实现,逻辑回归预测模型采用 Logit 变换,默认分布为二项分布。KNN 预测模型确定最近邻点数为 34。AdaBoost 预测模型采用二项分类方法,集成学习周期数为 7,弱分类器的模板类型选为‘Tree’。

为进一步验证机器学习识别异位性心搏的效果,将识别的结果按照下述规则计数为 FP,TP,TN 和 FN:

FP: 被识别为正常心搏的异位性心搏总数,即假阳性个数;

TP: 被正确识别为异位性心搏总数,即真阳性个数;

TN: 被识别为正常心搏总数,即真阴性个数;

FN: 被识别为异位性心搏的正常心搏总数,即假阴性个数。

在此基础上计算各类机器识别算法的曲线下面积(AUC)、准确度、灵敏度和特异性,所有的性能评价指标由接收者操作特征(Receiver Operating Characteristic, ROC)曲线计算得出。AUC 值代表 ROC 曲线下的面积。ROC 曲线是以灵敏度为纵坐标,1-特异性为横坐标绘制的曲线。该预测模型的灵敏度和特异性从 ROC 曲线的截断点处获得。

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FP}) \times 100\% \quad (2)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \times 100\% \quad (3)$$

采用五折交叉验证,由于阳性和阴性数据个数严重不对等,在交叉验证过程中,将数据分为阳性组和阴性组,每次训练、测试过程中,分别从 2 组中随机选取 4/5 的数据(阳性 12 586 个,阴性 162 590 个)作为训练集,余下 1/5 的数据(阳性 3 147 个,阴性 40 647 个)作为测试集,检验模型的预测能力。每得到一个特征组合,进行五折交叉验证,并重复 20 次求结果的均值,以确保算法的准确性。将数据集放入模型中进行特征筛选和交叉验证后,所得结果为最优特征组合以及相应的预测值和 ROC 曲线,从而得到 AUC、准确性、灵敏度和特异性的值。

2 实验结果

7 种机器学习算法对应的预测 ROC 曲线如图 3 所示。ROC 曲线能够从整体上反映一个分类器的预测性能,ROC 曲线下的面积即 AUC 取值范围为 0~1, AUC 值越大,表明分类器的预测性能越好。从图 3 可以看出,DT 和 AdaBoost 的 ROC 曲线向左上角(0,1)点突出明显, AUC 值更大,表明其预测性能优于 SVM,KNN,DT,QDA 和 AdaBoost,常用的 SVM 预测性能最低。与此同时,分析 7 种机器学习方法预测 AUC 值之间的统计性差异,通过计算每个分类器 AUC 的标准差和 Z 值,得到 7 种机器学习算法两两之间的单侧 p 值(如表 1 所示)。一般认为当 $p < 0.05$ 时,两者之间具有统计性差异。从表 1 可以看出,除 NB 和 QDA 之间无显著性差异外(从图 3(b)中也可以看出,二者之间具有非常相似的 ROC 曲线),其他分类器的预测结果均具有显著性区别,所有 $p < 0.014$ 。

ROC 曲线和 AUC 值仅能说明分类器的整体性能,为更好地比较各种分类器的预测性能,本文同时计算出分类器在 ROC 曲线截断点处的准确性、灵敏度和特异性值(如表 2 所示)。截断点就是 ROC 曲线最接近左上角(0,1)点所对应的坐标点,从而根据正确指数最大选出最佳临界点。正确指数又称约登指数(Youden's index),表示预测方法判断为真阳性与真阴性的总能力,是灵敏度与特异度之和减去 1,即约登指数=灵敏度+特异度-1。

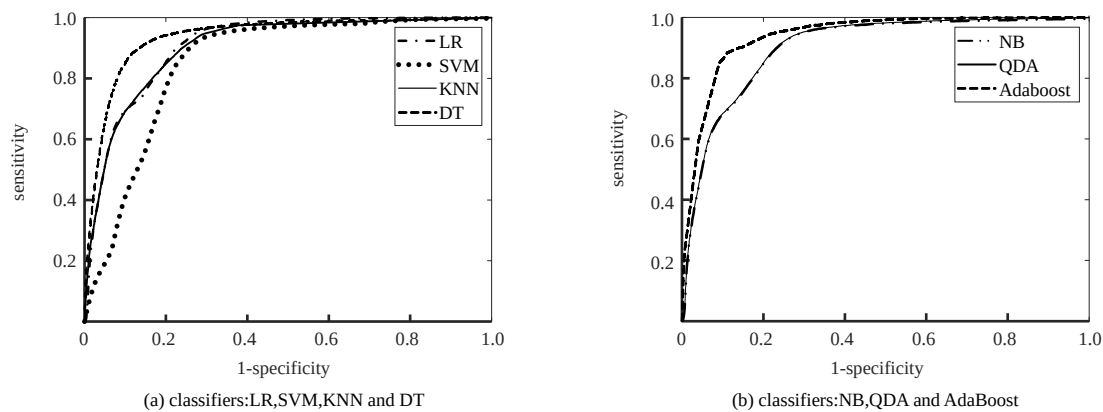


Fig.3 ROC curves of seven machine learning methods
图 3 七种机器学习算法对应的预测 ROC 曲线

表 1 七种机器学习方法预测 AUC 值之间两两比较的差异分析
Table1 Difference analysis of the AUC values predicted by seven machine learning methods in the pairwise comparison

<i>p</i>	LR	KNN	DT	NB	QDA
LR	0.500	<0.001	<0.001	<0.001	<0.001
KNN	<0.001	0.500	<0.001	0.007	0.001
DT	<0.001	<0.001	0.500	<0.001	<0.001
NB	<0.001	0.007	<0.001	0.500	0.273
QDA	<0.001	0.001	<0.001	0.273	0.500
AdaBoost	<0.001	<0.001	0.014	<0.001	<0.001
SVM	<0.001	<0.001	<0.001	<0.001	0.500

表 2 七种机器学习方法预测的 AUC 值、准确性、灵敏度和特异性值
Table2 AUC,accuracy,sensitivity and specificity values predicted by seven machine learning methods

machine learning method	AUC	accuracy	sensitivity	specificity
LR	0.910	0.752	0.940	0.737
KNN	0.904	0.771	0.899	0.761
DT	0.937	0.881	0.886	0.880
NB	0.900	0.763	0.917	0.752
QDA	0.899	0.764	0.914	0.753
AdaBoost	0.940	0.888	0.884	0.888
SVM	0.851	0.741	0.918	0.728

7 种机器学习算法各自的 AUC、准确性、灵敏度和特异性的值如表 2 所示。可以看出，集成学习 AdaBoost 的 AUC 值最高，为 0.940；同时准确度最高，为 0.888；且特异性最优，为 0.888，说明它的预测能力最强。集成学习 AdaBoost 分类器的预测能力最好，因为其能够对弱学习得到的弱分类器的错误进行适应性调整。灵敏度最高的是 LR，为 0.940，但对应的特异性最差，为 0.737。SVM 的预测能力最差，AUC 为 0.851，准确度为 0.741。一般 SVM 对小样本量的预测性能较好，因本文采用的数据为 10^5 量级，导致了 SVM 训练过程中的过拟合，预测性能不理想。

3 结论

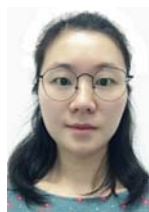
本文分析了 37 只 9~10 月龄小鼠(包括 12 只正常小鼠和 25 只心肌病小鼠)的 ECG 信号波形，计算了特征值——脉冲排除滤波和模板匹配的值，并经 3 位专家完成异位性心搏的标定及用于分类数据库的建立。分类时，使用 7 种机器学习方法组合特征值对异位性心搏点进行自动识别。实验结果表明，7 种机器学习方法都能得到较好的异位性心搏预测性能，即 AUC 值均大于 0.899，其中集成学习 AdaBoost 的预测能力最好，AUC 值可以达到 0.940，准确度、特异性均为 0.888。

参考文献：

[1] 刘彬. 心电信号特征识别及其在心血管疾病诊断中的应用[D]. 长春:吉林大学, 2014. (LIU Bin. Feature recognition of ECG signal and application in diagnosis of cardiovascular disease[D]. Changchun,Jilin,China:Jilin University, 2014.)
[2] FU D G. Cardiac arrhythmias:diagnosis,symptoms,and treatments[J]. Cell Biochem Biophys, 2015,73(2):291–296.
[3] 陈新. 临床心律失常学:电生理和治疗[M]. 北京:人民卫生出版社, 2000. (CHEN Xin. Clinical arrhythmia:electrophysiology and therapy[M]. Beijing:People's Medical Publishing House, 2000.)
[4] RAJENDRA A U,PAUL J K,KANNATHAL N,et al. Heart rate variability:a review[J]. Medical & Biological Engineering & Computing, 2006,44(12):1031–1051.
[5] SOLOSENKO A,MAROZAS V. Automatic extrasystole detection using photoplethysmographic signals[J]. IFMBE Proceedings, 2014,41(1):985–988.
[6] PALANIUPPAN R,KRISHNAN S M. Detection of ectopic heart beats using ECG and blood pressure signals[C]// International Conference on Signal Processing and Communications. Bangalore,India:IEEE, 2004:573–576.

- [7] VAIDYA G N. Application of exercise ECG stress test in the current high cost modern-era healthcare system[J]. Indian Heart Journal, 2017,69(4):551–555.
- [8] MCNAMES J, THONG T, ABOY M. Impulse rejection filter for artifact removal in spectral analysis of biomedical signals[C]// The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. San Francisco, USA: IEEE, 2004:145–148.
- [9] 李锋, 陈美丽. 一种基于几何特征的 ECG 波形识别算法[J]. 北京生物医学工程, 2015(3):261–266. (LI Feng, CHEN Meili. An ECG waveform recognition algorithm based on geometric characteristics[J]. Beijing Biomedical Engineering, 2015(3):261–266.
- [10] LIU C, LI L, ZHAO L, et al. A combination method of improved impulse rejection filter and template matching for identification of anomalous intervals in RR sequences[J]. Journal of Medical and Biological Engineering, 2012,32(4):245–250.
- [11] NGUYEN M, SU Y, KIRIAZIS H, et al. Upregulated galectin-3 is not a critical disease mediator of cardiomyopathy induced by β_2 -adrenoceptor overexpression[J]. American Journal of Physiology–Heart and Circulatory Physiology, 2018,314(6):H1169–H1178.
- [12] Nguyen M, KIRIAZIS H, RUGGIERO D, et al. Spontaneous ventricular tachyarrhythmias in β_2 -adrenoceptor transgenic mice in relation to cardiac interstitial fibrosis[J]. American Journal of Physiology–Heart and Circulatory Physiology, 2015,309(5):H946–H957.
- [13] HANIN L, ZAIDER M. Optimal schedules of fractionated radiation therapy by way of the greedy principle:biologically-based adaptive boosting[J]. Physics in Medicine Biology, 2014,59(15):4085–4098.
- [14] HUSSAIN H M, BENKRID K, SEKER H. Dynamic partial reconfiguration implementation of the SVM/KNN multi-classifier on FPGA for bioinformatics application[C]// 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC). Milan, Italia:[s.n.], 2015:7667–7670.
- [15] KANG J, SCHWARTZ R, FLICKINGER J, et al. Machine learning approaches for predicting radiation therapy outcomes:a clinician's perspective[J]. International Journal of Radiation Oncology, Biology, Physics, 2015,93(5):1127–1135.
- [16] 陈垚, 宋召青. 基于复合核支持向量回归机的多类分类算法[J]. 太赫兹科学与电子信息学报, 2017,15(6):1039–1044. (CHEN Yao, SONG Zhaoqing. Multi-class classification method based on support vector regression machine with composite kernel function[J]. Journal of Terahertz Science and Electronic Information Technology, 2017,15(6):1039–1044.)
- [17] OH E S, CHOI J H, LEE J W, et al. Predictors of intentional intoxication using decision tree modeling analysis:a retrospective study[J]. Clinical and Experimental Emergency Medicine, 2018,5(4):230–239.

作者简介:



何 密(1982–), 女, 四川省营山县人, 副教授, 主要研究方向为生物医学与雷达信号处理. email:hmcherry@126.com.

粘永健(1982–), 男, 山东省莱阳市人, 博士, 副教授, 主要研究方向为医学人工智能.

张 芸(1981–), 女, 长沙市人, 工程师, 主要研究方向为人工智能及应用.

胡厚源(1966–), 男, 湖北省荆州市人, 教授, 博士生导师, 主要研究方向为冠心病基础与临床.